Recent Deep Neural Networks for Object Detection

Jun Pan

Institute of International Education, School of Hebei University of Technology, Tianjin, China

Jun.pan@student.lut.fi

Abstract. Object recognition is a basic and difficult task in computer vision. Its purpose is to identify the object and give its exact position in the picture. In recent years, it has attracted extensive attention and gradually become a research hotspot. With the continuous development of object detection, some investigators have been solving existing problems and started to apply deep neural networks to the task. Despite the great progress in the research work surrounding the mission, there are few reviews of the mission, lacking a comprehensive review of its development in recent years. Given this period of rapid development, I will give an overview of object detection and present what methods have been used in object detection in recent years to improve task performance. This paper's goal is to give an exhaustive overview of the most latest developments in the area brought about by deep learning technology. At the same time, the possible problems in the task are analyzed, and the potential problems in the task are analyzed.

Keywords: Object detection, neural networks, deep learning.

1. Introduction

A long-standing, fundamental, and difficult issue in computer vision, object detection has been the topic of intensive investigation for many years. Developing computational models and methods for object recognition aims to answer one of the most fundamental questions needed for computer vision applications: what items are where? [1] The foundation of a lot of other things about computer vision, like instance segmentation, picture titling, object tracking, and so on, is object identification, one of the fundamental issues in computer vision. All these indicate the importance of object detection in computer vision. Object detection, the foundation of image understanding and computer vision, is to investigate techniques to detect many sorts of things in a coordinated skeleton to mimic human vision and cognition. Object detection is a basis for processing complex visual tasks. Examples of this are image captioning, segmentation, activity recognition, scene understanding, event detection, and object tracking. Object detection is useful in many applications. Examples include human-computer interaction, consumer electronics, autonomous driving, robot vision, security protection and so on. At present, the rapid development of deep learning technology provides impetus for the development of object detection. Solved a lot of problems that couldn't be solved before. So object detection has become a hot research topic.

The task initially mainly relied on machine learning techniques. In managed learning, the algorithm is given a data set and the correct response. The data is used by the machine to train it to calculate the right response. For a specific dataset, there is no "correct answer" in unsupervised learning; all data are identical. Supervised learning has two main tasks, one is regression, the other is classification. The task of unattended learning is to extract underlying structure from a given data set. Reinforcement learning focuses on how an agent acts in an environment to maximize returns. Reinforcement learning is expected to achieve higher intelligence because it is closer to the nature of biological learning. An agent can learn what behavior to take in what state by reinforcement learning. Reinforcement learning can be understood as learning by making mistakes. While deep learning requires a large amount of data and high requirements for hardware. Deep learning can solve problems, but it can't tell humans why. So these are some of the machine learning approaches. Firstly, the features of the task are constructed manually, and then some machine learning algorithms are used to classify or regression the task. However, there are two problems in this kind of method: first, the construction of task features depends on people's background knowledge, so the quality of features is unstable; Secondly, these features are shallow and mostly based on statistical methods, which cannot

well depict the deep semantic information. The emergence of deep learning solves the above problems well. Deep learning's capacity to infer new characteristics from a constrained set of features present in the training set gives it a significant advantage over earlier neural networks and other machine learning methods. Data scientists can save months of labor by relying on deep learning networks because these networks can generate features without being specifically instructed to do so. It also means that data scientists can use feature sets that are more complex than machine learning tools. Deep learning can automatically learn the relevant semantic information of a task through multi-layer nonlinear transformation.

Although deep neural network is applied to the task, it solves many problems and improves the performance of the task. However, there are very few reviews of this task. As a result, many beginners are unable to learn the latest progress of the task and what the latest methods have been proposed when they contact this task. In addition, the review can be regarded as a phased summary, which contributes to the development of subsequent research-based work. Therefore, I combed and summarized the development of this task in recent years. I first distinguished the definition of tasks, and then introduced the common data sets and evaluation indicators of tasks. In addition, I also explained the mainstream deep learning methods adopted by the task. Finally, I introduce some of the latest approaches to tasks and some of the potential challenges. Object detection can facilitate People's Daily life. Face recognition, for example, allows people to pay for items they want to buy using facial recognition alone. Moreover, object detection can also facilitate traffic management. It can not only complete traffic flow monitoring and traffic light timing control, improve traffic capacity. It can also detect and track traffic violations, greatly improving the supervision capacity of public security departments.

2. Background

2.1. Task Definition

A type of image segmentation called object detection, often called object extraction, because of geometric and statistical properties of objects. The combination of object segmentation and detection is the key factor for the conclusive and real-time enforcement of the system. The combination of object segmentation and detection is the key factor for the irrefutable and real-time implementing of the system. Automatic object extraction and detection is important, especially in complex scenarios where manifold elements need to be analyzed in real time. Because the computer principle is applied more and more widely and the computer technology is constantly updated, those who apply the computer image processing technology products are widely welcomed. Because it can track targets in real time.

2.2. Datasets

Throughout the history of object identification research, datasets have been crucial, and creating larger datasets with fewer biases is crucial for creating sophisticated computer vision algorithms. Datasets help the field move toward more difficult and complicated issues by providing a standard framework for evaluating and contrasting the performance of competing algorithms. Deep learning technology has recently had considerable success solving a variety of visual identification challenges, and the secret to this success is a big volume of labeled data.

VisDrone-DET2021 Dataset: Similar to the past three years, the Visdrone-DET2021 dataset uses the same dataset as the first three challenge datasets. Not only that, but target bounding boxes that exceed 540K are annotated by using 10 predefined categories. In the test set of this challenge, 1610 images from the Testdev subset will be used for public evaluation. An additional 1,580 images from the Test-Challenge subset will be used in the contest. Not only that, but also predefined 10 object categories such as person, car, van, truck, bicycle, motorcycle, tricycle, etc. On this basis, the results of AP, AP50, AP75, AR1, AR10, AR100 and AR50 detection algorithms can be evaluated. These algorithms can be analyzed using the evaluation protocol in MS COCO. Specifically, AP was

calculated by averaging over all 10 intersections of the Union (IoU) threshold (that is, a consistent step of 0.05 in the range of [0.50: 0.95]) for all categories, and it was used as the main indicator for ranking. AP50 and AP75 are calculated at single-IOU thresholds of 0.5 and 0.75 for all categories. The AR1 score is the average maximum number of recalls over all labels and IoU thresholds given by each image after one detection. The scores are the same and are the final scores after 10, 100 and 500 tests, respectively.

The dataset from the Visdrone-VDT2018 Challenge continues to be used in the Visdrone-Vid2019 Challenge, as shown in Figure 1. There are a total of 79 sequences in the data set, containing a total of 33,366 frames. There are three non-overlapping subsets, validation set, training set and test set in the data set. The training set includes 56 video clips with a total of 24,198 frames. There are seven video clips in the validation set with a total of 2846 frames. There are 16 video clips in the test set totaling 6,322 frames. These sequences were taken under different weather and light conditions and in different cities. AP, AP5u, AP75, AR1, AR10, AR100 and AR500 indexes were used for quantitative evaluation. AP50 and AP75 are the scores of a single IOU threshold of 0.5 and 0.75 for all object categories, respectively. ARi, AR10, AR100 and AR500 correspond to the maximum number of recalls of 1, 10, 100 and 500 detections per frame, respectively, which on average exceed all categories and IoU thresholds. [6]

Figure 1. Sample Images

2.3. Metrics

The question now is how to evaluate the effectiveness of the object detector? In fact, the answer to this query may change over time. There isn't a detection performance evaluation standard that is generally used in the early detection field. "Average Precision (AP)," which was first proposed in VOC2007, has become the most often used target detection evaluation metric in recent years. AP is normally calculated separately for each object category and evaluated in a category-specific manner. However, the average AP (mAP) of all object classes is typically used as the final performance indicator if you wish to compare the performance of all object classes. [1]

True Positive (TP): True indicates that the behavior of the classifier is correct. Positive indicates that the result given by the classifier is a Positive sample. In summary, TP represents a Positive sample predicted correctly by the classifier. True Negative (TN): True means that the behavior of the classifier is correct; Negative means that the result given by the classifier is a Negative sample. In general, TN means a Negative sample predicted correctly by the classifier. Accuracy is one of the most obvious indicators in the classification task, which is defined as the percentage of the total sample that correctly predicts the results. The so-called correct prediction refers to the original

positive sample, the classifier prediction result is also positive sample (TP); And if the original sample is negative, the prediction result of the classifier is also negative sample (TN).

$$\text{Accuracy} = (TP + TN)/(TP + TN + FP + FN) \quad (1)$$

When receiving the paper, we assume that the corresponding authors grant us the copyright to use the paper for the book or journal in question. When receiving the paper, we assume that the corresponding authors grant us the copyright to use.

3. Recent Researches

This paper points out that the problem of object detection is that the objects in the captured image are small in the distance and large in the near. And the object will be occluded in the dense environment, which makes the object detection more difficult. This article describes how to perform accurate object detection (accurately capturing objects) in an extremely dense urban scene. The solution to this problem is to mix the Anchor-free detector with the regression module to build the detector. In doing so, the model is released from the laborious process of leaping box size regression and performs better at multi-scale item detection in dense images. Two detectors are employed in place of an anchor frame to anticipate the center point and width of each target. These center points and diameters are then transformed into thick bounding boxes. The depth feature map and the coarse leaping box are then entered into a re-regression module. A thick border can be modified with the re-Regression module to create a precise, final border. In addition, an adaptive resampling enhancement strategy is introduced here to enhance the data logically. And that works really well. The problem that object detection may not be detectable in dense scenes is solved. This approach led to the team winning second place in the ICCV VisDrone2019 Image Object Detection Challenge. [3]

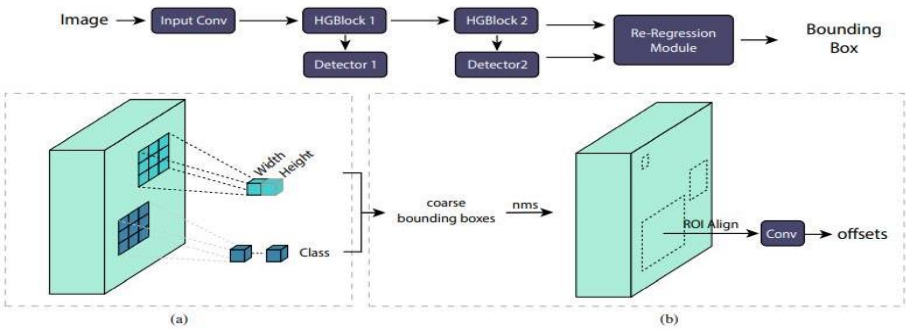

Figure 2. Model Architecture of RRNet

From the experimental results, it can be seen that the performance of the existing technology is inferior to RRNet. RRNet can handle very small objects in dense scenarios and is capable of meeting the highest performance of AP50, AR10 and AR100.

This paper points out that current object detection algorithms usually concentrate on the detection of ordinary scenes, rather than the scene captured by drones. The reason is that these studies are confined to a shortage of data sets. Therefore, Advancing the most advanced detection methods in drone scenarios, the 1st Visual and UAV Image Object Detection Challenge was held in combination with the 15th European Computer Vision Conference. This competition presents a larger scale UAV-based object detection dataset for evaluating detection algorithms in real-world scenarios. The authors proposed 34 target detection methods and evaluated their comprehensive performance. The competition uses algorithms submitted by researchers to detect 10 predefined objects in a data set, such as pedestrians and cars. From the analysis of 47 detectors on published datasets, 33 out of 47 detectors are better than the strong baseline cascaded RCNN detector.

This essay primarily addresses the issue of object detecting accuracy and speed. The authors developed a single-stage multi-scale framework to learn how to uniformly represent objects in different ranges. The authors call this UR3D. It can realize object detection quickly and accurately. Scale information can be used by UR3D to describe various detecting tasks. In this approach, the

demands on the model's capacity can be decreased, and it will be easier to accurately detect 3D things seen through a monocular. Additionally, by automatically choosing candidate boxes with superior distance estimates, distance bootstrap NMS improves distance estimation. The exact location of projected 2-D angles and 3-D frame centers is also proposed using an efficient fully convolutional cascaded point regression method, which may be able to recover the physical dimensions and orientation of objects. This results in a loss of projective projection consistency, enabling UR3D to achieve accurate monocular 3D object detection in compact architectures. [5]

This paper points out that although great progress has been made in object detection on still images, more and more attention has been paid to object detection in video. However, object detection still faces many challenges. Examples include drastic appearance changes, motion blurring and occlusion when extended from the image by state-of-the-art object detectors. From the data set collected by the drone, it can be known that the target detection and the drone are getting closer and closer. Such as camera variation and motion blur, algorithms for video object detection often cannot optimally process UAV-generated video sequences. To address this issue, the authors propose a more universal massive Visdrone-VDT2018 dataset to further advance the study of computer vision problems in UAV platforms. Doing so can promote the research of UAV video target detection.

The best performing detector in terms of AP scores, according to experiments with Visdron-Vid2019, is DBAI-DET (A.4), which combines a number of newly reported efficient top conference networks. Following DBAI-DET (A.4), similar promising results were obtained by AFSRNet (A.1) and HRDet+ (A.10), proving the value of multi-scale representation [6].

4. Challenges

So far, the first challenge the mission has encountered is the variety of perspectives. Because an object can look completely different from different angles. Therefore, the goal of the detector is to recognize objects from different angles. The second challenge is that there will be occlusion in object detection [7]. Sometimes, objects are obscured by other things, making it difficult to identify the objects' symbols. This situation adds to the difficulty of targeting. The third challenge is when things deform. The subject of computer vision analysis is not only a solid object, but can also deform and change its shape, which provides additional complexity for object detection [8-10]. The fourth challenge is light conditions. Lighting has a big influence on the definition of an object. The uniform object will look diverse depending on the lighting conditions. These are some of the challenges of object detection.

5. Conclusion

In computer vision, broad object detection is a significant and difficult subject that has drawn a lot of attention. Due to the incredible advancement of deep learning technology, the field of object detection has undergone enormous improvement. This paper's goal is to give an exhaustive overview of the most recent developments in this area brought about by deep learning technology. At the same time, the possible problems in the task are analyzed. The problems encountered in the process of developing goal detection in recent years are investigated in detail, and the corresponding good solutions to each problem are introduced in detail. In the future, it is hoped that object detection can be performed in any environment and anywhere with a high degree of accuracy.

References

[1] Zhengxia Zou, Zhenwei Shi, Member, IEEE, Yuhong Guo, and Jieping Ye, Senior Member, IEEE Object Detection in 20 Years: A Survey.

[2] Li Liu1, 2 Wanli Ouyang3 Xiaogang Wang4 Paul Fieguth5 Jie Chen2 Xinwang Liu1 Matti Pietikäinen Deep Learning for Generic Object Detection: A Survey.

[3] Changrui Chen, Yu Zhang, Qingxuan Lv, Shuo Wei, Xiaorui Wang, Xin Sun*, Junyu Dong Ocean University of China RRNet: A Hybrid Detector for Object Detection in Drone-captured Images.

[4] Pengfei Zhu 1, Dawei Du2, Longyin Wen 3, Xiao Bian 4, Haibin Ling 5.

[5] Qinghua Hu1 VisDrone-DET2019: The Vision Meets Drone Object Detection in Image Challenge Results

[6] Xuepeng Shi, Zhixiang Chen & Tae-Kyun Kim.

[7] Distance-Normalized Unified Representation for Monocular 3D Object Detection.

[8] Pengfei Zhu, Dawei Du, Longyin Wen, Xiao Bian, Haibin Lingo, Qinghua Hul, Tao Peng, Jiayu Zheng, Xinyao Wang, Yue Zhang, Liefeng Bo, Hailin Shi. VisDrone-VID2019: The Vision Meets Drone Object Detection in Video Challenge Results.

[9] Yaru Cao, Zhijian He2 Lujia Wang, Wenguan Wang, Yixuan Yuan, Dingwen Zhang, Jinglin Zhang, Pengfei Zhu, Luc Van Gool, Junwei Han, Steven Hoi VisDrone-DET2021: The Vision Meets Drone Object detection Challenge Results.

[10] Pal, S. K., Pramanik, A., Maiti, J., & Mitra, P. (2021). Deep learning in multi-object detection and tracking: state of the art. Applied Intelligence, 51(9), 6400-6429.